

Minimax wavelet estimation for multisample heteroscedastic non-parametric regression

Madison Giacomini[†], Sophie Lambert-Lacroix^{◦*}, Franck Picard^{*}

[†] Laboratoire LJK, Université de Grenoble et CNRS

[◦] UJF-Grenoble 1/CNRS/UPMF/TIMC-IMAG UMR 5525

^{*} LBBE, UMR CNRS 5558 Université Lyon 1

November 17, 2015

Abstract

The problem of estimating the baseline signal from multisample noisy curves is investigated. We consider the functional mixed effects model, and we suppose that the functional fixed effect belongs to the Besov class. This framework allows us to model curves that can exhibit strong irregularities, such as peaks or jumps for instance. The lower bound for the L_2 minimax risk is provided, as well as the upper bound of the minimax rate, that is derived by constructing a wavelet estimator for the functional fixed effect. Our work constitutes the first theoretical functional results in multisample non parametric regression. Our approach is illustrated on realistic simulated datasets as well as on experimental data.

keywords.-functional mixed effects models; minimax risk; Besov class; wavelet estimator;

*Corresponding author

1 Introduction

Functional data analysis has gained increased attention in the past years, in particular in high-throughput biology with the use of mass spectrometry. In this field, the signal is a spectrum whose peaks provide information regarding the protein content of biological samples. A new challenge in functional data analysis is the availability of multisample data for which functional ANOVA has become the appropriate framework. More specifically for spectrometry data, it is now well accepted that the noise corrupting the signal can be divided into a technical white noise added to an important inter-individual variability (Eckel-Passow et al., 2009). In this case, the usual non-parametric regression framework (a deterministic trend corrupted by a random noise) is no longer appropriate since it does not account for heteroscedastic noise structure. Functional mixed effects models (Antoniadis and Sapatinas, 2007) appear to be a powerful framework to handle these data, as others, and we focus here on the estimation of the baseline signal.

In practice, a trivial averaging procedure is often used to get an estimate of the baseline signal, but it has both a poor convergence rate and a finite sample performance. Amato and Sapatinas (2005) proposed an approach for baseline estimation based on empirical wavelet coefficients of the observed data. Unfortunately the convergence of their estimator is not theoretically assessed, and more broadly, there is a general lack of theoretical results on functional estimators in functional mixed models, despite their increasing importance in practice (Morris and Carroll, 2006; Morris et al., 2008).

In this work we propose a minimax estimator of the baseline signal, based on the empirical wavelet coefficients of the observed data. The functional fixed effect is assumed to belong to the Besov class, which allows us to model curves that can exhibit strong irregularities, such as peaks in mass spectrometry data. We construct the lower bound for the L_2 minimax risk. This convergence rate is the same as in the classical non parametric setting but with an additional approximation error term. Then, we propose a wavelet estimator that achieves near optimal rate of convergence (within a logarithmic factor in sample size). Through simulation studies, we show that our approach outperforms the approach proposed by Amato and Sapatinas (2005). We also propose a new thresholding procedure based on the Stein Unbiased Risk Estimate (SURE) (Stein, 1981), combined with the SCAD thresholding (Antoniadis and Fan, 2001). This leads to improved performance for the baseline signal estimation.

This article is organized as follows. Section 2 presents the heteroscedastic model and the theoretical properties of our minimax estimator (lower and upper

bounds). In particular we show how classical rates are modified in the presence of replicates along with inter-individual variability. Most of all, our work constitutes the first theoretical functional results in heteroscedastic multisample non-parametric regression. Several thresholding strategies are considered in Section 3, where we provide a new SURE-based procedure. Section 4 is devoted to the numerical experiments, and the procedure is illustrated on an experimental dataset. Technical proofs are provided in the Appendix.

2 Heteroscedastic nonparametric regression model and theoretical properties

2.1 Functional model

We observe N curves $Y_i(\cdot)$, for $i = 1, \dots, N$, over M equally spaced time points $\mathbf{t} = (t_1, \dots, t_M)$ in $[0, 1]^M$, with $M = 2^J$ for some integer J . In the general functional setting we consider a functional modeling (as in Antoniadis and Sapatinas (2007)) for the observed signal of the i th individual:

$$Y_i(t_j) = \mu(t_j) + E_i(t_j), \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, M, \quad (1)$$

where $E_i(\cdot)$, for $i = 1, \dots, N$, are stochastically independent random functions that are modeled as realizations of zero-mean Gaussian processes with parametrically structured covariances modeled in the wavelet domain (see Section 2.4). We define μ to be the main functional fixed effect characterizing a population average profile. In the following, we will denote by $\mathbf{Y}_i = (Y_i(t_1), \dots, Y_i(t_M))$, $i = 1, \dots, N$, the vector of observations on the time grid, and similarly by $\boldsymbol{\mu}$ and \mathbf{E}_i , $i = 1, \dots, N$, respectively the vector of the fixed effect and the noise terms, observed on the discrete time grid.

This modeling allows us to account for functional mixed effects models by decomposing $E_i(t)$ in a sum of two independent processes $E_i(t_j) = U_i(t_j) + \epsilon_{ij}$, where ϵ_{ij} are independent and identically distributed Gaussian random variables with zero-mean and constant variance; $U_i(t)$ is a centered Gaussian process standing for subject-specific functional deviations. In Amato and Sapatinas (2005), the authors introduce similar model although the variance of the process $U_i(t)$ is constant with respect to positions t_j .

2.2 Minimax approach

In what follows we suppose that μ belongs to the Besov class $\mathcal{F} = \mathcal{F}(s, p, q, L)$ (see Section 2.4 for a proper definition), a set of compactly supported functions (on $[0, 1]$) with a bounded Besov space norm (by L). Such a set allows to model curves that can exhibit strong irregularities, such as peaks or jumps for instance. The notion of regularity is at the core of the functional setting which makes inhomogeneous Besov spaces a privileged tool for irregular function analysis. These spaces allow the fine definition of the regularity s of a function along with its derivatives lying in $L^p([0, 1])$ while bringing a correction q to this regularity. For a detailed review of Besov spaces and their properties, we refer the reader to the books of Härdle et al. (1998) or DeVore and Lorentz (1993).

Our goal is to recover the main functional effect μ from noisy observations. An originality of our approach is to consider multiple, say N , individuals, which constitute available replicates to estimate the main fixed effect. To derive our estimator, we propose to use the so-called minimax approach. In this setup the risk of an estimator $\hat{\mu}_{N,M}$ is defined by $\mathbb{E}(\|\hat{\mu}_{N,M} - \mu\|)$, with $\|\cdot\|$ being a functional norm or a semi-norm. Then the so-called *minimax* estimator, denoted by $\hat{\mu}_{N,M}^*$, is the minimizer of the maximal risk on class \mathcal{F} over the set of all estimators:

$$\mathcal{R}(\hat{\mu}_{N,M}, \mathcal{F}) = \sup_{\mu \in \mathcal{F}} \mathbb{E}(\|\hat{\mu}_{N,M} - \mu\|).$$

Thus the challenge is to propose an optimal minimax estimator $\hat{\mu}_{N,M}^*$, and to derive its associated risk $\mathcal{R}_{N,M}^*(\mathcal{F}) = \mathcal{R}(\hat{\mu}_{N,M}^*, \mathcal{F})$, also referred to as the minimax risk.

The construction of minimax estimators on the Besov classes is well known when only one replicate is available (see Härdle et al. (1998)). When errors are measured with a L_r -norm “sharper” than the norm of the functional class p , wavelet-based thresholding estimators can significantly outperform linear projection estimates. The rate of convergence depends on r , p and s with two zones: the regular zone with usual rate $M^{-s/(2s+1)}$ and the sparse zone with a slower rate of convergence. However, this rate is not known when replicates are available ($N > 1$). In this work we establish this risk for $r = 2$ (we will denote this norm by $\|\cdot\|_2$) and for the Besov class \mathcal{F} with usual constraints $p \geq 1$, $q \geq 1$ and $s \geq 1/p$. That leads to consider the regular zone since, in this case, we have $s' = s - 1/p + 1/2 > 0$ (see Härdle et al. (1998)). In order to establish the minimax risk, we first give its lower bound and secondly we propose an estimator that achieves a near optimal rate of convergence. In this context, the near-optimality means that the minimax rate is attained within a logarithmic factor in sample size M .

2.3 Lower bound for the minimax risk

One of the main contributions of this paper is to derive the asymptotic lower and upper bounds for $\mathcal{R}_{N,M}(\mathcal{F})$. The following theorem gives the lower bound for this minimax risk in the inhomogeneous Besov class when dealing with multisample datasets (*i.e.* $N > 1$).

Theorem 2.1 *Under the model (1) with finite variances for the processes $E_i(\cdot)$, for $i = 1, \dots, N$, assume that μ belongs to a Besov class $\mathcal{F}(s, p, q, L)$ with $p \geq 1$, $q \geq 1$, $s \geq 1/p$ and $L < \infty$, then*

$$\mathcal{R}_{N,M}(\mathcal{F}) \geq \mathcal{O} \left[(MN)^{\frac{-s}{2s+1}} \right] + \mathcal{O} \left[M^{-s'} \right].$$

where $s' = s - 1/p + 1/2 > 0$, if $p < 2$, $s' = s$ otherwise.

Let us mention that the term $(MN)^{\frac{-s}{2s+1}}$ could be expected since it is the minimax rate (when $N = 1$) considering a noise of variance $\text{Var}(E_i(t_j))/N$. However the approximation error term $M^{-s'}$, present in the case with only one sample ($N = 1$), is always negligible compared with the term $(MN)^{\frac{-s}{2s+1}}$. When $N > 1$, even a large N does not provide more information on the function μ outside the grid (t_1, \dots, t_M) . Hence, $M^{-s'}$ becomes a limiting term.

2.4 Wavelet estimator of the functional effect

The upper bound of the minimax rate given in Theorem 2.1 is derived by constructing a wavelet estimator $\hat{\mu}_{N,M}$ of μ . Owing to their strong connection with the class of Besov spaces, wavelets indeed represent a powerful tool to perform adaptive functional regression (see Donoho et al. (1995)).

As a brief recall and to set notations, wavelets can be used to construct orthonormal basis of the functional Hilbert space $L^2([0, 1])$ by dilating and translating a compactly supported scaling function denoted by ϕ and a compactly supported mother wavelet denoted by ψ . We assume that ϕ and ψ belongs to $C^m([0, 1])$. Then, letting $j' \in \mathbb{N}$ be the first level of approximation, the family:

$$\{\phi_{j'k}, k = 0, \dots, 2^{j'} - 1; \psi_{jk}, j \geq j_0, k = 0, \dots, 2^j - 1\},$$

with $\phi_{j'k}(t) = 2^{j'/2} \phi(2^{j'}t - k)$ and $\psi_{jk}(t) = 2^{j/2} \psi(2^jt - k)$ is an orthonormal basis of $L^2([0, 1])$. Thus, any function μ in the space $L^2([0, 1])$ can be expressed

in the wavelet basis as:

$$\mu(t) = \sum_{k=0}^{2^{j'}-1} \alpha_{j'k}^* \phi_{j'k}(t) + \sum_{j \geq j'} \sum_{k=0}^{2^j-1} \beta_{jk}^* \psi_{jk}(t),$$

where $\alpha_{j'k}^* = \langle \mu, \phi_{j'k} \rangle$ and $\beta_{jk}^* = \langle \mu, \psi_{jk} \rangle$ are respectively the *theoretical* approximation and wavelet coefficients, and with $\langle \cdot, \cdot \rangle$ being the canonical Hilbertian scalar product associated with the space $L^2([0, 1])$. In the following, we set $j' = 0$ and omit the index $(0, 0)$ for the unique remaining scaling coefficient denoted by α^* .

The Besov class $\mathcal{F}(s, p, q, L)$ is defined via wavelet coefficients in the following way:

$$\mathcal{F}(s, p, q, L) = \{ \mu \in L^2([0, 1]) : \|\mu\|_{spq} \leq L \},$$

where

$$\|\mu\|_{spq} = |\alpha^*| + \left(\sum_{j=0}^{\infty} (2^{j(s-1/p+1/2)} \|\beta_{j\cdot}^*\|_p)^q \right)^{\frac{1}{q}}, \quad \|\beta_{j\cdot}^*\|_p = \left(\sum_{k=0}^{2^j-1} (\beta_{jk}^*)^p \right)^{\frac{1}{p}}.$$

For $p, q > 0$ and $1/p - 1 < s < m$, the norm $\|\cdot\|_{spq}$ is equivalent to the norm of the corresponding Besov space (cf. Donoho (1994), Delyon and Juditsky (1997)).

In statistical settings, we are more concerned with discretely sampled curves. By applying the fast discrete wavelet transform proposed by Mallat (1989) to the functional model (1), we obtain a representation of the model in the coefficient domain given by:

$$M^{-\frac{1}{2}} \mathbf{W} \mathbf{Y}_i = M^{-\frac{1}{2}} \mathbf{W} \boldsymbol{\mu} + M^{-\frac{1}{2}} \mathbf{E}_i, \quad \forall i = 1, \dots, N$$

$$\begin{bmatrix} c_i \\ \mathbf{d}_i \end{bmatrix} = \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_i^c \\ \boldsymbol{\varepsilon}_i^d \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} \boldsymbol{\varepsilon}_i^c \\ \boldsymbol{\varepsilon}_i^d \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}). \quad (2)$$

The $M \times 1$ vector $(c_i, \mathbf{d}_i^T)^T$ contains *empirical* scaling and wavelet coefficients associated with the signal, while $(\alpha, \boldsymbol{\beta}^T)^T$ stand for empirical coefficients related to the fixed effect μ and $(\boldsymbol{\varepsilon}_i^c, \boldsymbol{\varepsilon}_i^{d^T})^T$ for the coefficients coming from the error term \mathbf{E}_i . Following Antoniadis and Sapatinas (2007), the modeling of such correlated noise is performed directly in the wavelet domain by assuming first that \mathbf{G} is a diagonal matrix thanks to the well known decorrelating property of wavelets (see Frazier et al. (1991)). Then, to attain a wide range of processes, variances are assumed to vary with respect to the position and the resolution level such that

$Var(\varepsilon_i^c) = \sigma_c^2/\sqrt{M}$ and $Var(\varepsilon_{ijk}^d) = \sigma_{jk}^2/\sqrt{M}$ for all (j, k) in Λ with $\Lambda = \{(j, k) | j = 0, \dots, J-1; k = 0, \dots, 2^j-1\}$. Conversely, existing works dealing with a correlated noise focused on the modeling of individual noise processes in the time domain by assuming a stationnary noise (Johnstone and Silverman (1997)) or a locally stationnary noise (von Sachs and MacGibbon (2000)). In the wavelet domain these assumptions translate into variance terms for the matrix \mathbf{G} that are respectively depending on j (σ_j^2) or depending on both j and k . Based on the decorrelating property of wavelets, extra diagonal terms in the matrix \mathbf{G} are then neglected which restricts the class of reached processes in a way that is not effectively controlled. As a matter of fact, our model allows to consider non stationary processes whose covariance is diagonalizable by the DWT. However, we claim that such a modeling enables to catch a wide range of processes, even non stationary and hence allows a flexible enough modeling.

In the context of inhomogeneous spaces of functions such as Besov classes, it is known that in some cases, no linear method can achieve the optimal rate (see *e.g.* Härdle et al. (1998)) whereas nonlinear wavelet thresholding, pioneeringly introduced by Donoho and Johnstone (1994) in the white noise model, achieves this goal for a wide class of functional classes by taking advantage of the natural spatial adaptivity of wavelets. Starting from model (2) in the coefficient domain, we extend the usual thresholding procedures to the heteroscedastic framework by including position-dependent variance parameters in the thresholding expressions. For $N = 1$, the wavelet coefficients d_{1jk} are shrunk as from a certain level, through a defined shrinkage function δ , such that $\hat{\beta}_{jk} = \delta(d_{1jk}, \lambda_{jk})$, where $\lambda_{jk} = \lambda \hat{\sigma}_{jk}$, and λ is a regularization parameter to be fixed. The shrunk coefficients are inversely transformed to yield the solution in the time domain, namely $M^{\frac{1}{2}} \mathbf{W}^T [\hat{\alpha}, \hat{\beta}^T]^T$, where \mathbf{W}^T is the transpose of the orthogonal matrix \mathbf{W} .

When $N > 1$, Amato and Sapatinas (2005) propose three strategies to estimate μ in model (1) in the homoscedastic case.

1. The most natural one, widely used in practice, is the direct pointwise averaging of observations $\mathbf{Y}_1, \dots, \mathbf{Y}_N$. However this simple procedure leads to poor convergence rate as pointed by Amato and Sapatinas (2005), reflected by the completely pointwise procedure and poor finite sample performance. This approach is referred as a simple pointwise *average* approach by the authors.
2. The second approach consists in averaging the nonparametric regression curves of the N signals and is referred as a *shrink then average* approach.

This procedure improves the convergence rate due to the presence of a smoothing step.

3. The former strategy can be further improved by first averaging the observations $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ and then apply shrinkage to the average signal using then the whole sample. That is the third approach proposed in Amato and Sapatinas (2005) and referred as a *average then shrink* approach. Let us note that it has not been demonstrated that such an estimator achieves the optimal convergence rate.

In this work we consider this third approach in the heteroscedastic case and show that the associated estimator is near-minimax. Precisely, we consider

$$\hat{\boldsymbol{\mu}}_{N,M} = M^{\frac{1}{2}} \mathbf{W} \begin{bmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}, \quad (3)$$

with,

$$\begin{aligned} \hat{\alpha} &= c_{\bullet}, \\ \hat{\boldsymbol{\beta}} &= \begin{cases} d_{\bullet,jk}, & \text{for } j < j_0, \\ \delta(d_{\bullet,jk}, \lambda_{jk}), & \text{for } j = j_0, \dots, J-1, \end{cases} \end{aligned} \quad (4)$$

where \bullet denotes the average over the N samples. The choice of the parameter j_0 will be detailed in the proof of Theorem 2.2. The values of position-dependent thresholds λ_{jk} are then given by:

$$\lambda_{jk} = \hat{\sigma}_{jk} \frac{\sqrt{2 \log M}}{\sqrt{M}}, \quad (5)$$

where $\hat{\sigma}_{jk}^2$ are \sqrt{N} -consistent estimates of variances. The following result gives an upper bound for the quadratic risk depending on the signal size M and the number of samples N .

2.5 Upper bound of the minimax risk for wavelet-based thresholding estimators

Theorem 2.2 *Under the model (1), assume that μ belongs to a Besov class $\mathcal{F}(s, p, q, L)$ with $p \geq 1$, $q \geq 1$, $s \geq 1/p$, $L < \infty$, and that the variances σ_c^2 and $(\sigma_{jk}^2)_{(j,k) \in \Lambda}$ are*

bounded by a constant denoted by σ_{max}^2 . For any shrinkage function that satisfies, for any β and ξ ,

$$|\delta(\beta + \xi, \lambda) - \beta| < C(\min(|\beta|, \lambda) + |\xi| \mathbf{1}_{|\xi| > \lambda/2}), \quad (6)$$

then the estimator $\hat{\mu}_{N,M}$ defined by (3,4,5) with thresholds $2\lambda_{jk}$, satisfies

$$\mathbb{E}(\|\hat{\mu}_{N,M} - \mu\|_2) \leq \begin{cases} \max \left\{ \mathcal{O} \left[\left(\frac{\log M}{MN} \right)^{\frac{s}{2s+1}} \right] + \left[\mathcal{O} \left(\frac{\log M}{M} \right)^{s'} \right] \right\}, & \text{if } \frac{2}{2s+1} < p < 2 \\ \max \left\{ \mathcal{O} \left[\left(\frac{1}{MN} \right)^{\frac{s}{2s+1}} \right] + \left[\mathcal{O} \left(\frac{\log M}{M} \right)^{s'} \right] \right\}, & \text{if } p \geq 2 \end{cases}$$

where s' is defined as in Theorem 2.1.

The next section describes the practical derivation of thresholding procedures that satisfy the conditions required by Theorem 2.2. Thus we propose estimators that enjoy a near-optimal convergence rate in a multisample heteroscedastic setting.

3 Thresholding strategies

3.1 Shrinkage functions

Among the usual thresholding procedures we first focus on the hard and soft thresholding procedures of Donoho and Johnstone (1994), that provide estimators $\hat{\beta}^h$ and $\hat{\beta}^s$. We also consider the SCAD ($\hat{\beta}^{scad}$) thresholding of Antoniadis and Fan (2001) that establishes a trade-off between hard and soft thresholding, overcoming their respective non-continuity and bias drawbacks. The main conclusion of Theorem 2.2 is subject to the fulfilling of constraint (6). The Lemma 2 of Juditsky and Delyon (1996) ensures that hard and soft thresholding meet this requirement. Moreover, since we have

$$\forall \beta \in \mathbb{R}, \forall \lambda > 0, \quad \delta^s(\beta, \lambda) \leq \delta^{scad}(\beta, \lambda) \leq \delta^h(\beta, \lambda),$$

the conclusion of this Lemma still holds for the SCAD thresholding.

3.2 Choice of the threshold

For theoretical purposes, only the universal threshold (5) has been considered so far. Its easy implementation and its good asymptotic properties makes the universal threshold very popular in major wavelet packages. Our heteroscedastic thresholding approach is based on the definition of a threshold depending on the position (j, k) through the variance parameters (see (5)). Theorem 2.2 then applies in the context where the variances are unknown but for which \sqrt{N} -consistent estimates are available. When $N = 1$, exhibiting \sqrt{N} -consistent variance estimates is challenging. Such an issue has been considered in the litterature and approaches based on a functional modeling of the variances in the time domain have been developed (see *e.g.* Gasser et al. (1989), Antoniadis and Lavergne (1995), Cai and Wang (2008)). In their approaches, variances are then estimated using ν -order differences ($\nu \in \mathbb{N}$), coupled with an appropriate smoothing nonparametric method.

In the mutlisample context ($N > 1$), variance parameters can be easily estimated by simply considering empirical variances estimators such that:

$$\hat{\sigma}_{jk}^2 = \frac{1}{N-1} \sum_{i=1}^N (d_{ijk} - d_{\bullet jk})^2, \quad \text{for all } (j, k) \in \Lambda. \quad (7)$$

These variance parameter estimates straightforwardly satisfy the \sqrt{N} -consistency requirement due to their asymptotic normality properties.

However as pointed by Donoho and Johnstone (1994) and Coifman and Donoho (1995) the universal threshold, originally designed for a "noise-free" reconstruction, is substantially larger than the minimax threshold. To handle this practical drawback, Donoho and Johnstone (1995) proposed a strategy based on the Stein Unbiased Risk Estimate (SURE, Stein (1981)) whose purpose is to fix level dependent thresholds $\lambda_{\text{SURE},j}$ that leads to obtain an unbiased estimate of the L^2 -risk. Let $\tilde{\mathbf{d}}$ be a vector in \mathbb{R}^ℓ distributed as a standardized Gaussian distribution of mean β and covariance matrix equal to identity. The idea consists in writing the thresholding estimator $\hat{\beta}(\cdot) = \delta(\cdot, \lambda)$ as the sum:

$$\hat{\beta}(\tilde{\mathbf{d}}) = \tilde{\mathbf{d}} + \mathbf{g}(\tilde{\mathbf{d}}),$$

where \mathbf{g} is a weakly differentiable function from \mathbb{R}^ℓ to \mathbb{R}^ℓ . This leads to:

$$\mathbb{E} \left(\|\hat{\beta}(\tilde{\mathbf{d}}) - \beta\|_2^2 \right) = \ell + \mathbb{E} \left(\|\mathbf{g}(\tilde{\mathbf{d}})\|_2^2 + 2 \sum_{k=1}^{\ell} \frac{\partial \mathbf{g}(\tilde{\mathbf{d}})}{\partial d_k} \right).$$

The goal is then to select the parameter λ which minimizes the estimate of the L^2 -risk, denoted by $\text{SURE}(\lambda; \tilde{\mathbf{d}})$ and given by

$$\text{SURE}(\lambda; \tilde{\mathbf{d}}) = \ell + \|\mathbf{g}(\tilde{\mathbf{d}})\|_2^2 + 2 \sum_{k=1}^{\ell} \frac{\partial \mathbf{g}(\tilde{\mathbf{d}})}{\partial d_k}.$$

By considering $\tilde{d}_{jk} = d_{\bullet, jk} / \hat{\sigma}_{jk}$, where $\hat{\sigma}_{jk}^2$ is given as in (7), the SURE threshold is given by:

$$\lambda_{\text{SURE}, j} = \arg \min_{0 \leq \lambda \leq \lambda_{U, j}} \text{SURE}(\lambda, \tilde{\mathbf{d}}_j) \quad \text{for all } j = j_0, \dots, J-1, \quad (8)$$

where $\lambda_{U, j}$ is the universal threshold given in (5) and $M = 2^j$. The computation of the SURE criterion depends on the chosen thresholding function. Following the example of Donoho and Johnstone (1995) for soft thresholding, we propose an adaptation of the SURE concept to SCAD thresholding. Let us note that Park (2010) proposed an other derivation of the SURE criterion leading to a SURE-Block-SCAD estimator in the context of wavelet-based functional regression. When replicates are available, the SURE criterion to minimize according to λ is given by:

$$\begin{aligned} \text{SURE}_{\text{SCAD}}(\lambda; \tilde{\mathbf{d}}_j) &= 2^j + \sum_{k=0}^{2^j-1} (\tilde{d}_{jk}^2 - 2) \mathbf{1}_{\{|\tilde{d}_{jk}| \leq \lambda\}} + \sum_{k=0}^{2^j-1} \lambda^2 \mathbf{1}_{\{\lambda < |\tilde{d}_{jk}| \leq 2\lambda\}} \\ &+ \frac{1}{(a-2)^2} \sum_{k=0}^{2^j-1} \left[2(a-2) + \tilde{d}_{jk}^2 + (a\lambda)^2 + 2a\lambda |\tilde{d}_{jk}| \right] \mathbf{1}_{\{2\lambda < |\tilde{d}_{jk}| \leq \lambda\}}. \end{aligned} \quad (9)$$

The computation details can be found in Appendix 5.3. As recommended by Fan and Li (2001), a is set to 3.7 based on a Bayesian argument.

Moreover, we can point out that extremely sparse settings can lead to insufficient denoising due to the impact of zero coefficients in SURE criterion. To avoid this drawback, Donoho and Johnstone (1995) propose a compromising Hybrid Scheme (HS) between regular and SURE thresholding defined by:

$$\lambda_j^{\text{HS}} = \begin{cases} \lambda_{U, j} & \text{if } \sum_{k=0}^{2^j-1} d_{\bullet, jk}^2 \leq \hat{\sigma}_{jk}^2 2^{j/2} (2^{j/2} + j^{3/2}), \\ \lambda_{\text{SURE}, j} & \text{otherwise,} \end{cases} \quad (10)$$

for all $j = j_0, \dots, J-1$.

4 Numerical experiments

In this simulation study, we first investigate the benefits of using heteroscedastic thresholding estimators over homoscedastic ones when more than one sample are available. Then, we investigate the effect of the choice of the threshold on realistic simulated datasets.

Simulation settings. We consider the test functions `Blocks`, `Bumps`, `Heavisine` and `Doppler` (Donoho and Johnstone, 1994) that we use to model the principal mean functions μ . These functions are processed with the Daubechies' extremal phase wavelet basis with respectively 1, 2, 5 and 7 vanishing moments, based on the Shannon entropy as described in Nason (2008), Chap 2. Then we get the noise-free wavelet coefficients, to which we add heteroscedastic noise, following model (2): multisamples are simulated in the wavelet domain by corrupting the wavelet coefficients of the mean function by a normally additive heteroscedastic noise whose variance σ_{jk}^2 at a given position (j, k) in Λ is given by:

$$\sigma_{jk}^2 = \begin{cases} \sigma^2 & \text{for } (j, k) \in \Lambda_1, \\ \sigma^2 + 2^{-j\eta} \gamma_{jk}^2 & \text{for } (j, k) \in \Lambda_0. \end{cases} \quad (11)$$

The set $\Lambda_0 \subset \Lambda$ contains index associated with the zero coefficients of the mean function whereas Λ_1 contains the ones associated with nonzero coefficients. The first term σ^2 is associated to a white noise added to all coefficients, whereas the second term is an extra variability that introduces heteroscedasticity at some positions. Following Antoniadis and Sapatinas (2007), a scale-wise exponential decrease is imposed to the extra variability terms by the quantity $2^{-j\eta}$. Parameter η relates to the fixed effect regularity allowing the extra variability associated to γ_{jk}^2 to remain interpretable. In the following we use $\eta = 1.5$.

Dealing with zero and non-zero coefficients. One expects heteroscedastic thresholding estimators to be favored by heteroscedasticity structure expressed on the zero coefficients of the mean function: true zero coefficients are indeed more susceptible to be thresholded in this setting since heteroscedastic thresholds are expected to be larger than the homoscedastic one. Therefore we put emphasis on configurations where heteroscedasticity concerns the null wavelet coefficients of the mean function.

The value of σ^2 is controlled by a Signal-to-Noise Ratio (SNR) and takes values in (1,5) going from a high level (SNR=1) to a low level of noise (SNR=5). Parameters γ_{jk}^2 are then drawn from a Gamma distribution with scale 2 and shape $\gamma_{\text{ref}}^2/2$. The quantity γ_{ref}^2 associated to the heteroscedasticity intensity is controlled with respect to the baseline variance σ^2 by a ratio parameter τ defined by

$$\tau = \frac{M\sigma^2}{\gamma_{\text{ref}}^2 \sum_{(j,k) \in \Lambda_0} 2^{-j\eta}}.$$

Parameters values. We set the signal size to $M = 2048$ and the sample size to $N = 100$. A wider simulation study (not shown for the sake of clarity) reveals that the main conclusions do not differ with different signal and sample size. For each fixed effect function, the simulation design explores the following configurations: $\text{SNR} \in (1, 5)$, $\tau \in (0.1, 1)$. The variability and heteroscedasticity parameters σ^2 and γ_{ref}^2 are deduced from the value of SNR and τ respectively. Each configuration is repeated 200 times.

Heteroscedastic versus homoscedastic thresholding. We start by considering the framework defined by the assumptions of Theorem 2.2, *i.e.* we consider the SCAD thresholding function with the universal threshold in a heteroscedastic setting. Since the threshold used in Theorem 2.2 is known to be large (Donoho and Johnstone, 1994), it is set to half of its value in the following. Then heteroscedastic thresholding (denoted He) refers to the procedure that uses empirical estimates of the variance at each position $(j, k) \in \Lambda$ whereas homoscedastic thresholding (denoted Ho) uses $\hat{\sigma}_{MAD}^2$ (based on the Median Absolute Deviation (MAD) of the coefficients at the finest resolution level (Donoho and Johnstone, 1994)). Amato and Sapatinas (2005) introduced the idea of wavelet-based thresholding in the context of noisy repeated measurements and discussed how to integrate the replicates in the analysis. They use in (5) the usual robust variance estimate $\hat{\sigma}_{MAD}^2$ instead of position-dependent estimators $\hat{\sigma}_{j,k}^2$. However, they do not investigate the effect of the choice of the threshold, and they do not handle the potential heteroscedasticity in their synthetic data, despite the presence of inter-individual variability. A simulation study (not shown) revealed that the strategy of taking the mean of the individual MAD leads to better performance. Therefore we consider this strategy for the homoscedastic part. We aim at comparing homoscedastic and heteroscedastic procedures regarding to the mean function reconstruction performance. Performance of compared procedures are evaluated with respect to the Mean Integrated Squared Error (MISE) of the reconstructed mean function.

4.1 Results.

Average MISEs are presented on Table 1. The results show that heteroscedastic estimates greatly outperform homoscedastic ones in terms of functional reconstruction for all considered configurations. As expected, this is especially true when the heteroscedasticity intensity is high (*i.e.* for $\tau = 0.1$).

Table 1 here

Another argument supporting the use of heteroscedastic thresholding procedures concerns their adaptative behaviour in an homoscedastic framework: indeed, a simulation study in the homoscedastic framework (*i.e.* with $\sigma_{jk}^2 = \sigma^2$ for all $(j, k) \in \Lambda$) reveals similar reconstruction properties of homoscedastic and heteroscedastic estimates for a SCAD thresholding using the universal threshold. Corresponding results are displayed in Table 2.

Table 2 here

Comparing heteroscedastic procedures Despite good asymptotic properties, using the universal threshold may not be optimal in finite dimensional setting as mentioned by Donoho and Johnstone (1994) in their original paper. Therefore we now focus on comparing heteroscedastic procedures for different choices of thresholds on simulated datasets. In order to consider more realistic cases, we consider datasets where heteroscedasticity corrupts both null and non null coefficients of the mean function. Hence, starting from the same mean functions, the heteroscedasticity is as from now defined such that for (j, k) in Λ :

$$\sigma_{jk}^2 = \sigma^2 + \pi_{jk} \times 2^{-j\eta} \gamma_{jk}^2. \quad (12)$$

The quantities σ_{jk}^2 and γ_{jk}^2 are defined as previously whereas π_{jk} is assumed to be a realization of a Bernoulli distribution with parameters 0.3. Note that the pairs fixed effects- μ /heteroscedasticity structure- $\boldsymbol{\pi} = (\pi_{jk})_{(j,k) \in \Lambda}$ are kept fixed for all the synthetic datasets.

For each mean function associated to a given heteroscedastic structure $\boldsymbol{\pi}$, the simulation design explores the following configurations: SNR varies in $(1, 5)$ and τ in $(0.1, 0.25, 1)$. Similarly, the signal and sample size are respectively set to $M = 2048$ and $N = 100$ whereas each configuration is repeated 200 times. Examples of simulated data are represented on Figure 1 for all considered main patterns.

Figure 1 here

Then heteroscedastic thresholding procedures are compared for both Soft and SCAD thresholding functions, δ^s and δ^{scad} , and for both Universal and SURE threshold, λ_U and λ_{HS} . Performance of the procedures are evaluated with respect to the Mean Integrated Squared Error (MISE) of the reconstructed mean functions. Simulation results are presented on Figure 2. Examples of reconstruction associated to median performance are represented in Figure 3

Figure 2 and Figure 3 here

As a main conclusion we can observe that using the SURE threshold leads to improved performance for the reconstruction of the main effect in a heteroscedastic setting. As mentioned by Donoho and Johnstone (1995) in the homoscedastic framework, the universal threshold turns out to be too large in practice when dealing with finite dimensional signals.

Another interesting point concerns the interaction between the choice of the threshold and the thresholding function. When using the universal threshold, the SCAD thresholding gives indeed at least similar or improved reconstruction performance. This is expected since the SCAD thresholding is designed to smoothly correct the bias on high coefficients introduced by the soft thresholding. Conversely such a difference vanishes when using the SURE threshold for which Soft and SCAD thresholdings exhibit similar performance. This finding can be explained by the adaptive behaviour of the SURE threshold that compensates the existing bias on high coefficients.

By way of conclusion, the overall simulation study encourages the use of the heteroscedastic thresholding in the context of functional regression with multiple samples. Heteroscedastic thresholding keeps indeed the simplicity and the computational efficiency of the usual homoscedastic thresholding while being able to handle potential inter-individual variations. Moreover, in practice, using the adaptive SURE threshold, paired with the SCAD thresholding which enjoys good theoretical properties leads to improved reconstruction of the mean function.

As a last remark, we shall mention that the wider simulation study abovementioned with various sample and signal sizes shows that the overall MISEs orders of magnitude are more improved by a higher number of samples N than by a larger signal size M .

4.2 Analysis of experimental data

As an application to the proposed methodology, we analysed a SELDI-TOF mass spectrometry dataset issued from a study on ovarian cancer (Petricoin et al., 2002). This dataset was produced by the Ciphergen WCX2 protein chip and is publicly available through the Clinical Proteomics Programs Databank (¹, ovarian dataset 8-7-02). The sample set consists of 162 serums profiles from women affected by an ovarian cancer and 91 control subjects. Each spectra contains the measure of 15154 intensities characterizing as many mass over charge (m/z) ratios. Prior to analysis, raw data are background corrected using a quantile regression procedure, and spectra are aligned using a procedure based on wavelets zero crossings (Antoniadis et al., 2007). Moreover, we restrict on 512 intensities for m/z ratios within the range [5200,5915] centered around the main central peak. Mass spectrometry data represented a meaningful application for our method since Giacomini et al. (2013) show evidence for the presence of inter-individual variations occurring at specific ranges of m/z ratios resulting in a sharp heteroscedasticity structure.

We separately analysed the control group and the group affected by a cancer using an heteroscedastic SCAD thresholding procedure, with a SURE threshold. Mean reconstructed functions superimposed on experimental data are represented in Figure 4.

Figure 4 here

We can observe that individuals from the control and cancer groups exhibit similar mean functional profiles. Such an observation indicates that a nonparametric testing procedure would be on purpose to ascertain the presence of a significant effect of the group. Although it is out of the scope of the present paper, in this context, taking into account the presence of potential inter-individual variations appears as critical for the application of such testing procedure.

Acknowledgements

Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24. We are grateful to Anatoli Juditsky for constructive and fruitful discussions.

¹<http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>

References

- Amato, U. and T. Sapatinas (2005). Wavelet shrinkage approaches to baseline signal estimation from repeated noisy measurements. *Advances and Applications in Statistics* 51, 21–50.
- Antoniadis, A., J. Bigot, S. Lambert-Lacroix, and F. Letue (2007). Non parametric pre-processing methods and inference tools for analyzing time-of-flight mass spectrometry data. *Current Analytical Chemistry* 3(2), 127–147.
- Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association* 96(455), 939–955.
- Antoniadis, A. and C. Lavergne (1995). Variance function estimation in regression by wavelet methods. *Lecture Notes in Statistics "Wavelets and Statistics"* 103, 31–42.
- Antoniadis, A. and T. Sapatinas (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis* 51(10), 4793–4813.
- Cai, T. and L. Wang (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics* 36(5), 2025–2054.
- Coifman, R. and D. Donoho (1995). Translation invariant de-noising. *Wavelets and statistics* 103, 125–150.
- Delyon, B. and A. Juditsky (1997). On the computation of wavelet coefficients. *J. Approximation Theory* 88(1), 47–79.
- DeVore, R. and G. Lorentz (1993). *Constructive approximation*. Springer Verlag.
- Donoho, D. (1994). *Smooth wavelet decompositions with blocky coefficient kernels*. L. Schumaker, ed., ‘Recent advances in wavelet analysis’.
- Donoho, D. and I. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81(3), 425–455.
- Donoho, D. and I. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* 90, 1200–1224.

- Donoho, D., I. Johnstone, G. Kerkycharian, and D. Picard (1995). Wavelet shrinkage: asymptopia. *Journal of the Royal Statistical Society, Ser. B* 57(2), 371–394.
- Eckel-Passow, J. E., A. L. Oberg, T. M. Therneau, and H. R. Bergen (2009, Jul). An insight into high-resolution mass-spectrometry data. *Biostatistics* 10, 481–500.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.
- Frazier, M., B. Jawerth, and G. Weiss (1991). *Littlewood-Paley Theory and the Study of function Spaces*. Number 79. American Mathematical Society.
- Gasser, T., L. Stroka, and C. Jennen-Steinmetz (1989). Residual variance and residual pattern in nonlinear regression. *Biometrika* 73, 625–633.
- Giacofci, M., S. Lambert-Lacroix, G. Marot, and F. Picard (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* 69(1), 31–40.
- Härdle, W., G. Kerkycharian, D. Picard, and A. Tsybakov (1998). *Wavelets, Approximation and Statistical Applications*. Springer.
- Johnstone, I. and B. Silverman (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Ser. B* 59(2), 319–351.
- Juditsky, A. and B. Delyon (1996). On minimax wavelets estimators. *Applied and computational harmonic analysis* 3, 215–228.
- Mallat, S. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Transactions of the American Mathematical Society* 315(1), 69–87.
- Morris, J. S., P. J. Brown, R. C. Herrick, K. A. Baggerly, and K. R. Coombes (2008, Jun). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* 64, 479–489.
- Morris, J. S. and R. J. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B Stat Methodol* 68, 179–199.
- Nason, G. (2008). *Wavelet methods in Statistics with R*. Springer, New York.

- Park, C. (2010). Block thresholding wavelet regression using $\{\text{SCAD}\}$ penalty. *Journal of Statistical Planning and Inference* 140(9), 2755 – 2770.
- Petricoin, E. F., A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta (2002, Feb). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359, 572–577.
- Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* 9(6), 1135–1151.
- von Sachs, R. and B. MacGibbon (2000). Nonparametric curve estimation by wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statistics* 27(3), 475–499.

5 Appendix

5.1 Proof of Theorem 2.1

First let us recall the aim of the proof concerning the lower bound in the minimax context. Since

$$\mathbb{E}(\mathcal{R}_{N,M}^{-1}(\mathcal{F}) \|\hat{\mu}_{N,M} - \mu\|_2) \geq c_1 \mathbb{P}(\|\hat{\mu}_{N,M} - \mu\|_2 \geq c_1 \mathcal{R}_{N,M}(\mathcal{F})),$$

for some $c_1 > 0$, we have to show that

$$\mathbb{P}(\|\hat{\mu}_{N,M} - \mu\|_2 \geq c_1 \mathcal{R}_{N,M}(\mathcal{F})) > c_2,$$

for some constant $c_2 > 0$. Next we reduce the class \mathcal{F} to a subclass \mathcal{F}_n of finite number n of functions in \mathcal{F} because the *sup* is greater over a larger class. The family $\mathcal{F}_n = \{\mu_0, \dots, \mu_{n-1}\}$ is constructed by small perturbation of μ_0 , so that the distance between each pairs of functions is small and at least of order $\mathcal{R}_{N,M}(\mathcal{F})$. Then the problem can be reduced to the one of testing by the following way

$$\sup_{\mu \in \mathcal{F}_n} \mathbb{P}(\|\hat{\mu}_{N,M} - \mu\|_2 \geq c_1 \mathcal{R}_{N,M}(\mathcal{F})) \geq p_n = \inf_{\phi} \max_{j=0, \dots, n-1} \pi_{\phi}(\mu = \mu_j),$$

with π_{ϕ} the power function associated to ϕ , where ϕ is any test that allows to distinguishing between the n hypotheses, the k -th of them stating that the observations of model (1) are drawn from the k -th element of the set \mathcal{F}_n . To bound p_n

by $c_2 > 0$, we need to major the maximum of the Kullback distance $K(\mu_i, \mu_j)$ between observations of model (1) associated with μ_i and the ones associated with μ_j . For instance when $n = 2$, we have

$$p_2 \geq \max \left(\frac{\exp(-K(\mu_1, \mu_0))}{4}, \frac{1 - \sqrt{\frac{K(\mu_1, \mu_0)}{2}}}{2} \right).$$

Without loss of generality, since variances are assumed to be bounded, we can consider model (1) with $E_i(t_j)$, $i = 1, \dots, N$, $j = 1, \dots, M$, independent and identically distributed Gaussian random variables with zero-mean and variance σ_E^2 . In this case, we have

$$K(\mu_1, \mu_0) = \frac{N}{2\sigma_E^2} \sum_{j=1}^M (\mu_1(t_j) - \mu_0(t_j))^2. \quad (13)$$

Let us come back to the proof of the lower bound. This proof can be decomposed in two steps. For the usual term in $\mathcal{O} \left[(MN)^{\frac{-s}{2s+1}} \right]$, we just have to use the usual proof for the Besov classes by adding the factor N because of the multiplicative term N in (13). We now give the proof corresponding to the term in $\mathcal{O} [M^{-s'}]$. We only need two functions in order to construct \mathcal{F}_n , that is $n = 2$. For $p \geq 2$, we put $\mu_0(t) = 0$, for all $t \in [0, 1]$, and

$$\mu_1(t) = M^{\frac{1}{p}-s} \eta \left(Mt - \frac{1}{2} \right),$$

where $\eta \in \mathcal{F}(s, p, q, L)$ with support equal to $[-1/2, 1/2]$ such that $\eta(-1/2) = \eta(1/2) = 0$, and $\|\eta\|_2 \geq c > 0$. We have $\mu_0 \in \mathcal{F}(s, p, q, L)$ and

$$\|\mu_1\|_{spq} \leq M^{\frac{1}{p}-s} M^{s-\frac{1}{p}} \|\eta\|_{spq} \leq L.$$

So we also have $\mu_1 \in \mathcal{F}(s, p, q, L)$, and

$$\|\mu_1 - \mu_0\|_2 = M^{-\frac{1}{2}} M^{\frac{1}{p}-s} \|\eta\|_2 \geq M^{-s'} c,$$

hence, the family \mathcal{F}_2 is included in the Besov class $\mathcal{F}(s, p, q, L)$ and the L_2 -distance between the two functions are at least $M^{-s'}$. Since

$$K(\mu_1, \mu_0) = \frac{N}{2\sigma_E^2} M^{\frac{2}{p}-2s} \eta^2 \left(Mt_1 - \frac{1}{2} \right) = \frac{N}{2\sigma_E^2} M^{\frac{2}{p}-2s} \eta^2 \left(\frac{1}{2} \right) = 0,$$

we have

$$p_2 \geq 1/2,$$

and

$$\mathbb{E}(\|\widehat{\mu}_{N,M} - \mu\|_2) \geq \frac{c}{2}M^{-s'},$$

for any estimator $\widehat{\mu}_{N,M}$.

For $p \geq 2$, we use the same method but by choosing

$$\mu_1(t) = M^{-s} \sum_{j=1}^M \eta \left(Mt - j + \frac{1}{2} \right).$$

So we have

$$\begin{aligned} \|\mu_1\|_{spq} &\leq M^{-s} M^{\frac{1}{p}} M^{s-\frac{1}{p}} \|\eta\|_{spq} \leq L, \\ \|\mu_1 - \mu_0\|_2 &= M^{\frac{1}{2}} M^{-\frac{1}{2}} M^{-s} \|\eta\|_2 \geq M^{-s} c, \end{aligned}$$

and

$$K(\mu_1, \mu_0) = \frac{N}{2\sigma_E^2} M^{-2s} \sum_{j=1}^M \sum_{k=1}^M \eta \left(Mt_j - k + \frac{1}{2} \right) = \frac{N}{2\sigma_E^2} M^{-2s+1} \eta \left(\frac{1}{2} \right) = 0,$$

which concludes the proof.

5.2 Proof of Theorem 2.2

This proof is an adaptation of the proof of Theorem 1 of Juditsky and Delyon (1996). We denote by $\widetilde{\beta}_{jk}$, the estimator $\widehat{\beta}_{jk}$ with σ_{jk} instead of $\widehat{\sigma}_{jk}$ in (5) and $\widetilde{\mu}_{N,M}$ the associated estimator. We introduce

$$\mu_{j_1}(t) = \alpha \phi_{00}(t) + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk} \psi_{jk}(t),$$

where j_1 is such that $M/\log M \leq 2^{j_1} \leq 2M/\log M$. Let us note that (see proposition 1 of Delyon and Juditsky (1997)), there exists some constant C_0 such that this function belongs to $\mathcal{F}(s, p, q, C_0 L)$. The global quadratic risk can then be

decomposed such that:

$$\begin{aligned}
\mathbb{E}(\|\widehat{\mu}_{N,M} - \mu\|_2^2) &\leq \mathbb{E}(\|\widehat{\mu}_{N,M} - \widetilde{\mu}_{N,M}\|_2^2) + \mathbb{E}(\|\widetilde{\mu}_{N,M} - \mu_{j_1}\|_2^2) + \|\mu - \mu_{j_1}\|_2^2 \\
&\leq \mathbb{E}\left[\sum_{j=j_0+1}^{j_1} \sum_k |\widehat{\beta}_{jk} - \widetilde{\beta}_{jk}|^2\right] \\
&\quad + \mathbb{E}(|\widehat{\alpha} - \alpha|^2) + \mathbb{E}\left[\sum_{j=0}^{j_0} \sum_k |\widetilde{\beta}_{jk} - \beta_{jk}|^2\right] \\
&\quad + \mathbb{E}\left[\sum_{j=j_0+1}^{j_1} \sum_k |\widetilde{\beta}_{jk} - \beta_{jk}|^2\right] + \|\mu - \mu_{j_1}\|_2^2 \\
&= T_1 + T_2 + T_3 + T_4 + T_5.
\end{aligned} \tag{14}$$

We seek to bound from above each term of the decomposition. By using the delta method based on a Taylor expansion of the thresholding function and since $\widehat{\sigma}_{jk}^2$ are \sqrt{N} -consistent estimates of variances, we get:

$$\begin{aligned}
T_1 &\leq \sum_{j=j_0+1}^{j_1} \sum_k C_1 \frac{2N-1}{(MN)^2} \sigma_{jk}^4 \\
&\leq C_1 \sigma_{\max}^4 \frac{2^{j_1}}{M^2 N} \leq C_1 \sigma_{\max}^4 \frac{(\log M)^{-1}}{MN},
\end{aligned}$$

with C_1 being a positive constant. The model (2) leads to

$$c_{\bullet} \sim \mathcal{N}\left[\alpha, \frac{\sigma_c^2}{NM}\right] \quad \text{and} \quad d_{\bullet,jk} \sim \mathcal{N}\left[\beta_{jk}, \frac{\sigma_{jk}^2}{MN}\right].$$

Approximation coefficients in T_2 are left unchanged, hence we have:

$$T_2 = \mathbb{E}(|c_{\bullet} - \alpha|^2) \leq \frac{\sigma_{\max}^2}{MN}.$$

In the same way, terms in T_3 are not thresholded, hence we get:

$$T_3 = \mathbb{E}\left[\sum_{j=0}^{j_0} \sum_k |\widetilde{\beta}_{jk} - \beta_{jk}|^2\right] = \sum_{j=0}^{j_0} \sum_k \mathbb{E}(|d_{\bullet,jk} - \beta_{jk}|^2) \leq C_3 2^{j_0} \frac{\sigma_{\max}^2}{NM},$$

with C_3 being a positive constant. The term T_5 is the approximation term that can be bounded such that (see proposition 1 of Delyon and Juditsky (1997)):

$$T_5 \leq C_5 2^{-2j_1 s'} \leq \left[\frac{\log M}{M}\right]^{2s'}.$$

Finally, bounding term T_4 needs the use of constraint (6) with $\lambda' = 2\lambda$, we have

$$T_4 \leq \underbrace{\mathbb{E} \left[\sum_{j=j_0+1}^{j_1} \sum_k \min \left(|\beta_{jk}|, \frac{\lambda' \sigma_{jk}}{\sqrt{MN}} \right)^2 \right]}_{T_{4.1}} + \underbrace{\mathbb{E} \left[\sum_{j=j_0+1}^{j_1} \sum_k |\varepsilon_{\bullet,jk}^d|^2 \mathbf{1}_{|\varepsilon_{\bullet,jk}^d| > \frac{\lambda' \sigma_{jk}}{2\sqrt{MN}}} \right]}_{T_{4.2}},$$

where $\varepsilon_{\bullet,jk}^d = d_{\bullet,jk} - \beta_{jk}$. For the term $T_{4.1}$, since $\mu_{j_1} \in \mathcal{F}(s, p, q, C_0 L)$, we obtain:

$$\begin{aligned} T_{4.1} &\leq \sum_{j=j_0+1}^{j_1} \sum_k \left(2\lambda \frac{\sigma_{\max}}{\sqrt{MN}} \right)^{2-p} |\beta_{jk}|^p \\ &\leq C_{4.1} \left(\frac{2 \log M}{M} \right)^{1-\frac{p}{2}} \left(\frac{\sigma_{\max}^2}{N} \right)^{1-\frac{p}{2}} \underbrace{\sum_{j=j_0+1}^{j_1} \sum_k |\beta_{jk}|^p}_{=\mathcal{O}(2^{-s'pj_0})} \\ &\leq C_{4.1} \left(\frac{\log M}{M} \right)^{1-\frac{p}{2}} \left(\frac{\sigma_{\max}^2}{N} \right)^{1-\frac{p}{2}} 2^{-s'pj_0}. \end{aligned}$$

For $T_{4.2}$, we have with Cauchy-Schwartz and exponential inequalities:

$$\begin{aligned} T_{4.2} &\leq \sum_{j=j_0+1}^{j_1} \sum_k 9 \mathbb{E} (|\varepsilon_{\bullet,jk}^d|^4)^{\frac{1}{2}} \mathbb{E} \left[\left(\mathbf{1}_{|\varepsilon_{\bullet,jk}^d| > \lambda \sigma_{jk} / \sqrt{MN}} \right)^2 \right]^{\frac{1}{2}} \\ C_{4.2} &\leq \sum_{j=j_0+1}^{j_1} \frac{\sigma_{\max}^2}{MN} \exp \left[\frac{- \left(\lambda \sigma_{jk} / \sqrt{MN} \right)^2}{2 \sigma_{jk}^2 / MN} \right]^{\frac{1}{2}} \\ &\leq C_{4.2} \frac{\sigma_{\max}^2}{N} M^{-2} 2^{j_1} \leq C_{4.2} \frac{\sigma_{\max}^2}{MN} (\log M)^{-1}. \end{aligned}$$

In order to fix the parameter j_0 , the terms T_3 and $T_{4.1}$ need to be balance according to M , which leads to:

$$2^{j_0} = \mathcal{O} \left[(\log M)^{\frac{1-p/2}{1+s'p}} (MN)^{\frac{p/2}{1+s'p}} \right].$$

By replacing 2^{j_0} in terms T_3 and $T_{4.1}$, the inequality (14) becomes:

$$\begin{aligned} \mathbb{E} (\|\hat{\mu}_{N,M} - \mu\|_2^2) &\leq C_1 \frac{(\log M)^{-1}}{MN} + \frac{\sigma_{\max}^2}{MN} + C_3 \sigma_{\max}^2 \left[\frac{\log M}{MN} \right]^{\frac{2s}{2s+1}} (\log M)^{\frac{-2s'}{2s+1}} \\ &\quad + C_{4.1} \sigma_{\max}^{2-p} \left[\frac{\log M}{MN} \right]^{\frac{2s}{2s+1}} (\log M)^{\frac{-2s'}{2s+1}} \\ &\quad + C_{4.2} \sigma_{\max}^2 \frac{M^{-\frac{1}{8}}}{N \log M} + C_5 \left[\frac{\log M}{M} \right]^{2s'} \end{aligned}$$

The convergence of the overall expression is limited by the terms in $\mathcal{O} \left[\left(\frac{\log M}{M} \right)^{2s'} \right]$ and in

$$\mathcal{O} \left[\left(\frac{\log M}{MN} \right)^{\frac{2s}{2s+1}} (\log M)^{\frac{-2s'}{2s+1}} \right].$$

The latter leads to a limitation in

$$\begin{aligned} &\mathcal{O} \left[\left(\frac{\log M}{MN} \right)^{\frac{2s}{2s+1}} \right] && \text{if } \frac{2}{2s+1} < p < 2 \\ &\mathcal{O} \left[\left(\frac{1}{MN} \right)^{\frac{2s}{2s+1}} \right] && \text{if } p \geq 2 \end{aligned}$$

Hence, we get:

$$\mathbb{E} (\|\hat{\mu}_{N,M} - \mu\|_2^2) \leq \max \left\{ \mathcal{O} \left[\left(\frac{\log M}{MN} \right)^{\frac{2s}{2s+1}} \right] + \left[\mathcal{O} \left(\frac{\log M}{M} \right)^{2s'} \right] \right\},$$

that concludes the proof.

5.3 Derivation of the SURE criterion for SCAD thresholding

For recall, the SCAD thresholding function is given by:

$$\delta^{\text{scad}}(d_{jk}, \lambda, a) = \begin{cases} \text{sign}(d_{jk})(|d_{jk}| - \lambda)_+ & \text{si } |d_{jk}| \leq 2\lambda, \\ \frac{(a-1)d_{jk} - a\lambda \text{sign}(d_{jk})}{a-2} & \text{si } 2\lambda < |d_{jk}| \leq a\lambda, \\ d_{jk} & \text{si } |d_{jk}| > a\lambda. \end{cases} \quad (15)$$

and we are looking for a function $\mathbf{g} : \mathbb{R}^{2j} \rightarrow \mathbb{R}^{2j}$ such that:

$$\delta^{\text{scad}}(\mathbf{d}_j, \lambda, a) = \mathbf{d}_j + \mathbf{g}(\mathbf{d}_j), \quad (16)$$

to define the SURE-SCAD criterion:

$$\text{SURE}_{\text{scad}}(\lambda; \mathbf{d}_j) = 2^j + \|\mathbf{g}(\mathbf{d}_j)\|_2^2 + 2 \sum_{k=0}^{2^j-1} \frac{\partial \mathbf{g}(\mathbf{d}_{jk})}{\partial d_{jk}}, \quad (17)$$

with $\mathbf{g}(\mathbf{d}_j) = (g_{j0}(d_{j0}), \dots, g_{2^j-1}(d_{j,2^j-1}))$. By defining g as the weakly differentiable function:

$$\begin{aligned} g(d_{jk}) = & -d_{jk} \mathbf{1}_{\{|d_{jk}| \leq \lambda\}} - \lambda \text{sign}(d_{jk}) \mathbf{1}_{\{\lambda < |d_{jk}| \leq 2\lambda\}} \\ & + \left(\frac{d_{jk}}{a-2} + \frac{a\lambda \text{sign}(d_{jk})}{a-2} \right) \mathbf{1}_{\{2\lambda < |d_{jk}| \leq a\lambda\}}, \end{aligned}$$

with $\mathbf{g}(\mathbf{d}_j) = (g(d_{j0}), \dots, g(d_{j,2^j-1}))$, the relation (16) is satisfied. We can then compute:

$$\begin{aligned} \|\mathbf{g}(\mathbf{d}_j)\|_2^2 &= \sum_{k=0}^{2^j-1} g(d_{jk})^2 \\ \text{with } g(d_{jk})^2 &= d_{jk}^2 \mathbf{1}_{\{|d_{jk}| \leq \lambda\}} + \lambda^2 \text{sign}(d_{jk})^2 \mathbf{1}_{\{\lambda < |d_{jk}| \leq 2\lambda\}} \\ &\quad + \frac{1}{(a-2)^2} [d_{jk}^2 + (a\lambda)^2 + 2a\lambda|d_{jk}|] \mathbf{1}_{\{2\lambda < |d_{jk}| \leq a\lambda\}} \\ \sum_{k=0}^{2^j-1} \frac{\partial g(d_{jk})}{\partial d_{jk}} &= \sum_{k=0}^{2^j-1} \left[-\mathbf{1}_{\{|d_{jk}| \leq \lambda\}} + \frac{1}{a-2} \mathbf{1}_{\{2\lambda < |d_{jk}| \leq a\lambda\}} \right], \end{aligned}$$

which leads finally to the criterion (9) in Section 3.

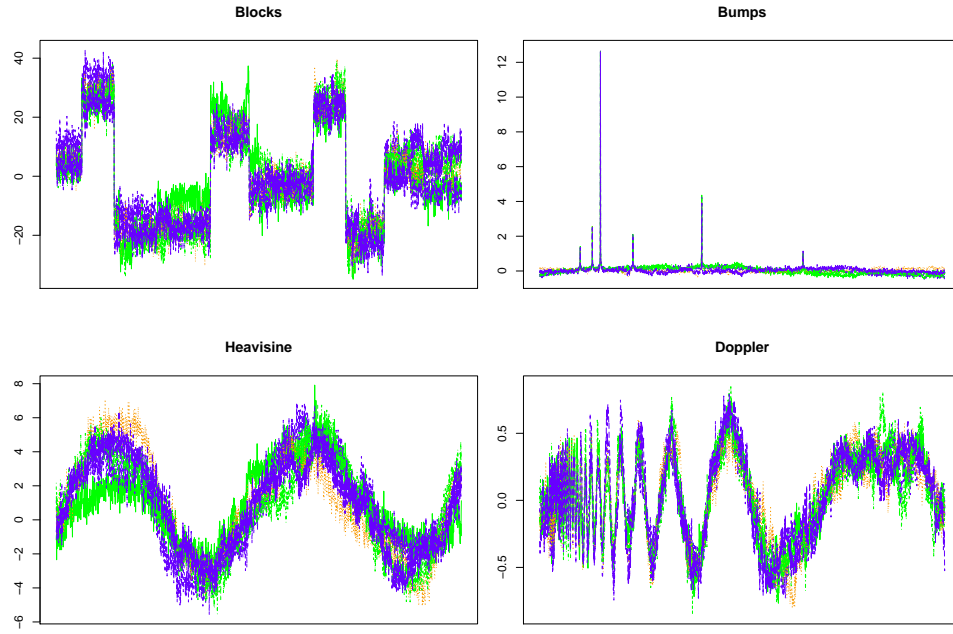


Figure 1: Examples of realistic simulated data. For each mean functions `Blocks`, `Bumps`, `Heavisine` and `Doppler`, 5 random realizations are represented. The parameters SNR and τ are respectively set to 5 and 0.25.

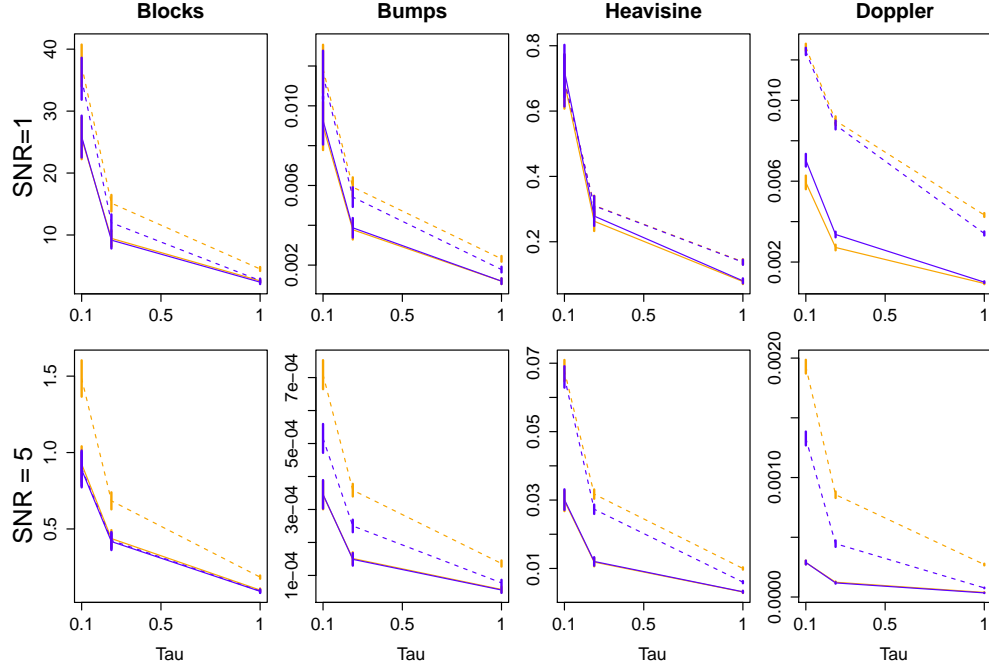


Figure 2: Resulting MISEs averaged over 200 repetitions for reconstructed fixed effects. Two SNR values in rows ($\text{SNR} = (1, 5)$ for a high/low noise) and three heteroscedasticity intensities on the horizontal axis of each graph ($\tau = 0.1, 0.25, 1$ from a high level to a low level) are considered. Soft and SCAD thresholding functions differ by plotting colors (respectively in orange and blue) whereas threshold choices Universal and SURE differ by the line types (respectively in dashed and solid line). Vertical bars are associated to the standard deviations of the resulting MISEs.

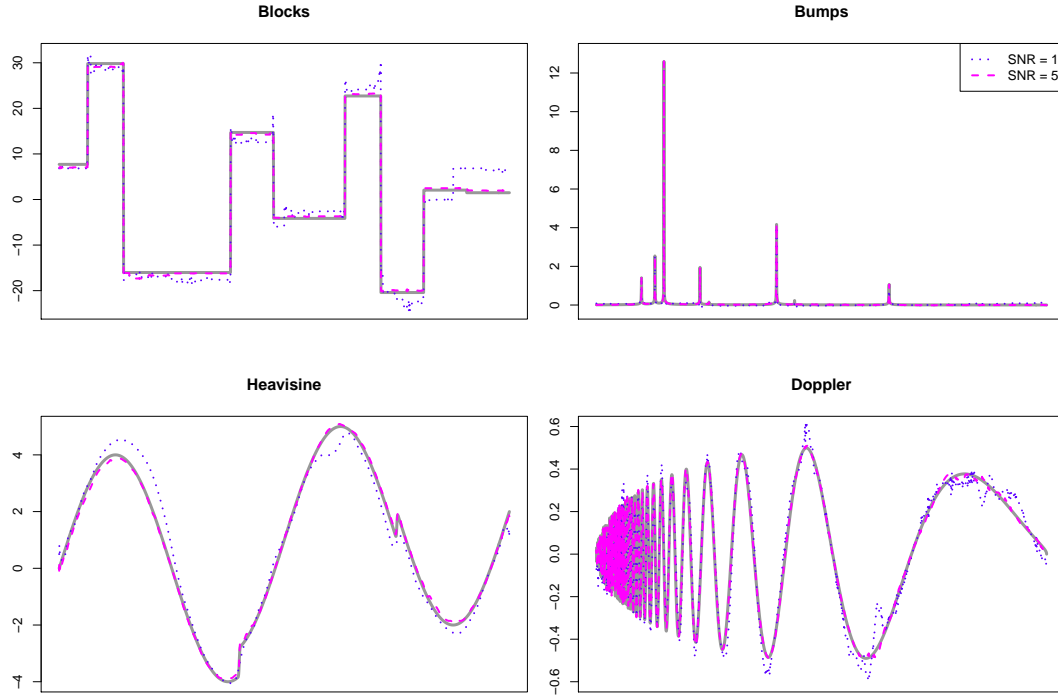


Figure 3: Examples of reconstructed mean functional effect using an heteroscedastic SCAD thresholding with the SURE threshold for models `Blocks`, `Bumps`, `Heavisine` and `Doppler`. The true mean functions is displayed in plain gray line. The parameter τ is equal to 0.25 whereas SNR take the values 1 (for a high noise, displayed in dotted blue lines) and 5 (for a low noise, displayed in dashed magenta lines). In all configurations, the chosen realization correspond to the one giving rise to the median MISE.

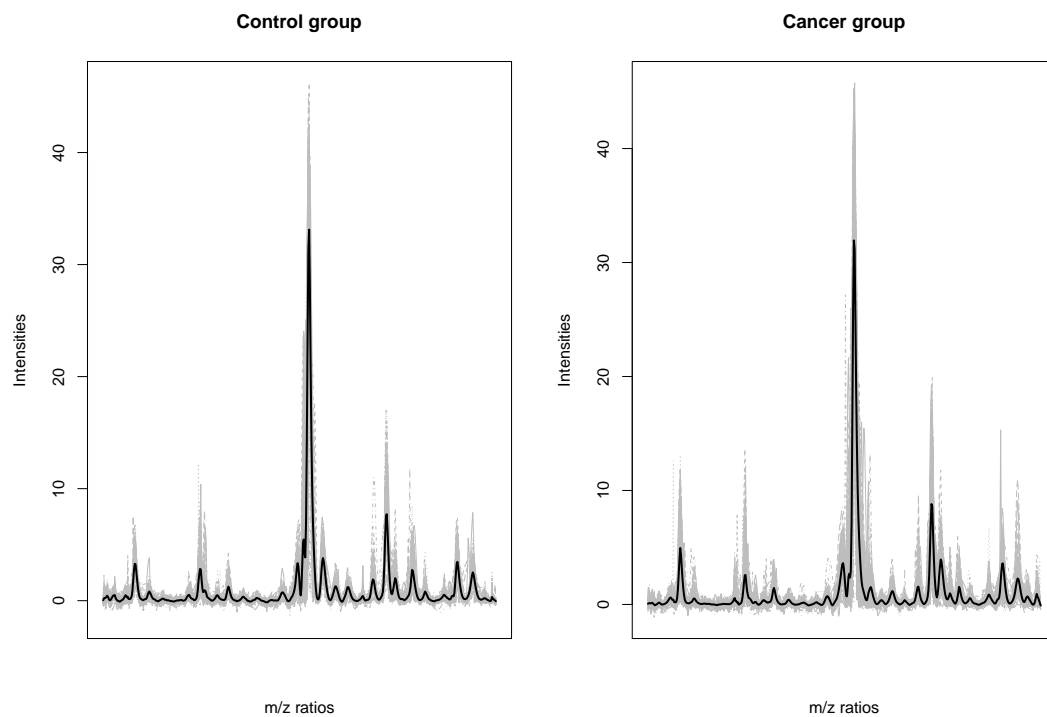


Figure 4: Mean reconstructed functions (bold line) superimposed on observed data (in light gray) for the control group and the cancer group.

	SNR = 1				SNR= 5			
	$\tau = 0.1$		$\tau = 1$		$\tau = 0.1$		$\tau = 1$	
	Ho	He	Ho	He	Ho	He	Ho	He
Blocks	5.093 (1.629)	0.168 (0.018)	0.424 (0.130)	0.166 (0.017)	0.186 (0.054)	0.001 (2e-4)	0.011 (0.006)	0.001 (2e-4)
Bumps	5.028 (0.745)	0.724 (0.025)	0.944 (0.048)	0.720 (0.027)	0.220 (0.029)	0.040 (0.001)	0.0573 (0.002)	0.040 (0.001)
Heavisine ($\times 10^{-2}$)	5.293 (0.303)	1.193 (0.103)	1.773 (0.120)	1.192 (0.104)	0.530 (0.016)	0.079 (0.006)	0.129 (0.008)	0.079 (0.006)
Doppler ($\times 10^{-4}$)	26.79 (4.058)	5.607 (2.607)	7.819 (0.320)	5.629 (0.238)	1.387 (0.136)	0.187 (0.117)	0.304 (0.015)	0.188 (0.010)

Table 1: Average MISE (and associated standard deviations) on 200 repetitions for the fixed effects Blocks, Bumps, Heavisine and Doppler in a heteroscedastic framework. The heteroscedastic structure is defined as in equation 11 with SNR and τ varying respectively in (1,5) and (0.1,1). The sample size is set to $N = 100$ and the signal size to $M = 1024$. Final estimates are based on a SCAD thresholding using the universal threshold λ_U .

	SNR = 1		SNR= 5	
	Homoscedastic	Heteroscedastic	Homoscedastic	Heteroscedastic
Blocks	0.189 (0.016)	0.168 (0.017)	1.44e-3 (2.5e-4)	1.43e-3 (2.5e-4)
Bumps	0.736 (0.024)	0.726 (0.024)	0.045 (1.25e-3)	0.040 (1.25e-3)
Heavisine ($\times 10^{-2}$)	1.203 (0.097)	1.204 (0.104)	0.079 (0.006)	0.078 (0.006)
Doppler ($\times 10^{-4}$)	5.658 (0.246)	5.622 (0.274)	0.201 (0.011)	0.188 (0.011)

Table 2: Average MISE (and associated standard deviations) on 200 repetitions for the fixed effects `Blocks`, `Bumps`, `Heavisine` and `Doppler` in a homoscedastic framework (with $\sigma_{jk}^2 = \sigma^2$ for all $(j, k) \in \Lambda$). The noise level is controlled by the SNR ratio varying in (1,5). The sample size is set to $N = 100$ and the signal size to $M = 1024$. Final estimates are based on a SCAD thresholding using the universal threshold λ_U .